# Data Mining Lab, Exercise 1

Team# 1: Poliakov Valerii, Holovnia Dmytro, Selvaraj Sinju

Dataset: data1.csv

## Task# 1

Calculate the mean and the standard deviation for the variable after excluding NA values. Draw the histogram.

```
> data<-read.csv(file="data1.csv", header=FALSE)
> col1<-data[,1]
> col1
  [1] 587.55 587.55 827.10  70.81  81.02 146.60 196.93 326.60 198.04     NA 513.66 590.78 682.03 677.49
 [15]     NA     NA     NA 213.80 563.18 339.67 818.65 728.44 600.89  62.16 287.87  92.14 265.59  32.64
 [29] 779.87 902.56 288.85 597.71 656.26 253.51 622.51  31.12  45.47 473.05  85.54 453.08 176.05 914.80
 [43] 145.48  32.96 231.88 352.15 524.77 595.33 871.21 333.87 210.49 874.48  69.56 775.50 540.42 510.54
 [57] 721.48 556.92 449.96 285.20 163.62 658.48     NA  71.72 181.48  55.79     NA     NA 630.76 581.88
 [71] 469.30 719.09     NA     NA     NA 244.36 391.24 371.79 764.12 356.25 300.92     NA     NA 439.63
 [85] 670.44     NA 189.93 670.17 639.93 142.47 142.47     NA     NA 852.21 203.36 476.41 162.73 366.60
 [99] 261.04  53.34 276.45
```

There is no column name in the .csv file and data contain NA values.

```
> clean_x<-na.omit(col1)
> clean_x
 [1] 587.55 587.55 827.10  70.81  81.02 146.60 196.93 326.60 198.04 513.66 590.78 682.03 677.49 213.80
[15] 563.18 339.67 818.65 728.44 600.89  62.16 287.87  92.14 265.59  32.64 779.87 902.56 288.85 597.71
[29] 656.26 253.51 622.51  31.12  45.47 473.05  85.54 453.08 176.05 914.80 145.48  32.96 231.88 352.15
[43] 524.77 595.33 871.21 333.87 210.49 874.48  69.56 775.50 540.42 510.54 721.48 556.92 449.96 285.20
[57] 163.62 658.48  71.72 181.48  55.79 630.76 581.88 469.30 719.09 244.36 391.24 371.79 764.12 356.25
[71] 300.92 439.63 670.44 189.93 670.17 639.93 142.47 142.47 852.21 203.36 476.41 162.73 366.60 261.04
[85]  53.34 276.45
attr(,"na.action")
 [1] 10 15 16 17 63 67 68 73 74 75 82 83 86 92 93
attr(,"class")
[1] "omit"
```

NA values excluded.

```
> mean(clean_x)
[1] 411.1831
```

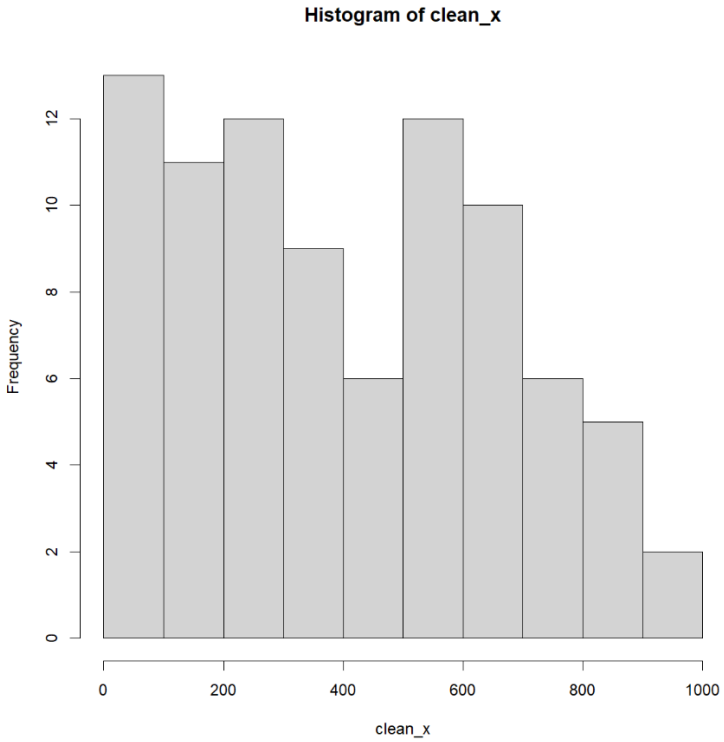
Mean after excluding NA values is 411.1831

```
> sd(clean_x)
[1] 257.6073
```

Standard deviation is 257.6073.

```
> hist(clean_x)
```

Histogram:




## Task# 2

Replace the missing values with the mean calculated in Task 1. Draw the histogram. Calculate the mean and the standard deviation after replacement.

```
> clean_y<-ifelse(is.na(col1), mean(col1, na.rm = TRUE), col1)
```

NA values replaced by the mean value.

```
> mean(clean_y)
```

```
[1] 411.1831
```

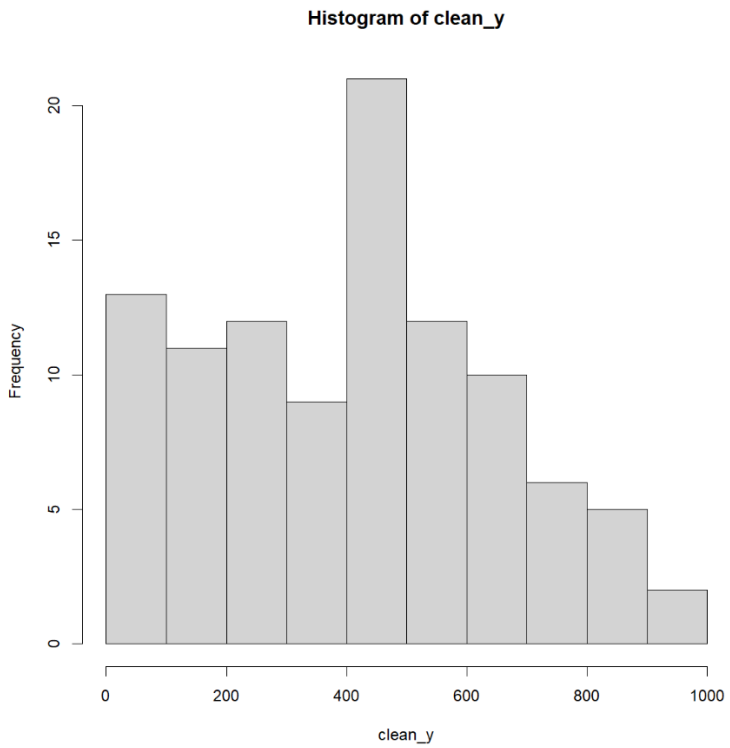
Mean value is 411.1831.

```
> sd(clean_y)
[1] 237.5022
```

Standard deviation is 237.5022.

```
> hist(clean_y)
```

Historgam:



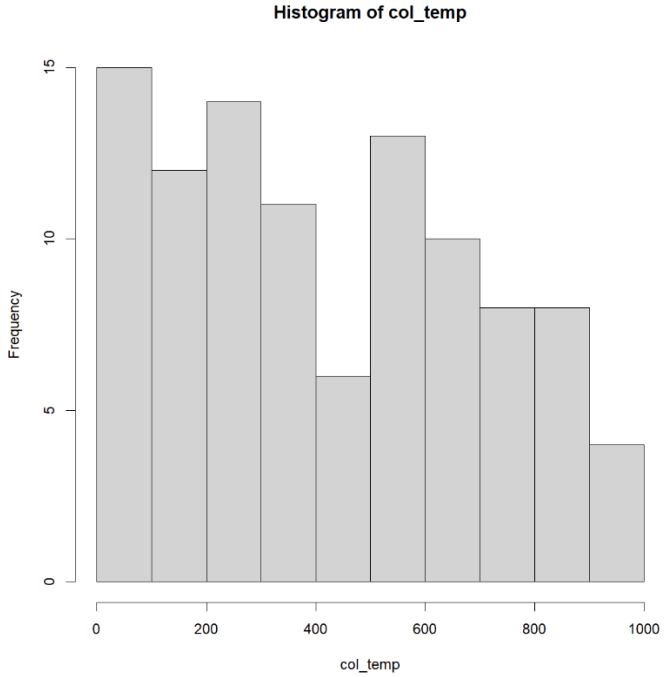
## Task# 3

Replace the missing values with the values generated at random from the observed variable distribution. Draw the histogram. Calculate the mean and the standard deviation after replacement.

```
> col_temp<-col1
> for(i in seq_along(col_temp)) {
+   col_temp[i]<-ifelse(is.na(col1[i]), sample(na.omit(col1), 1), col1[i])
+ }
> mean(col_temp)
[1] 428.9259
> sd(col_temp)
[1] 271.8829
> hist(col_temp)
```

After code execution, the NA values replaced with the random not NA value. Mean value is 428.9259, Standard Deviation is 271.8829.

Histogram:



Task# 4

Compare the results of Tasks 1, 2 and 3. Comment on the values of the mean and standard deviation. Compare the shapes of the histograms.

| | NA excluded | NA replaced with mean | NA replaced with random |
|---|---|---|---|
| Mean value | 411.1831 | 411.1831 | 428.9259 |
| Standard Deviation | 257.6073 | 237.5022 | 271.8829 |

Comparing histograms we see that replacing NA with the random values from the dataset we obtain distribution of values very close to the original dataset. By replacing NA with mean we receive significant increase of values equal to mean which might interfere further dataset analysis.

By replacing NA with mean values, the mean value of the new dataset will be the same as original dataset and standard deviation (SD) will be smaller than SD in the original dataset.

By replacing NA with random values from the original dataset, the mean values will be slightly differed from the original dataset and SD will be greater than SD in the original dataset.